STATISTIQUE, TECHNOLOGIE ET CULTURE

Defays
Eurostat, L-2920 Luxembourg

1. Introduction

La statistique n'a jamais été indépendante de son époque, de sa culture, de sa technologie. Elle s'apparente en cela aux autres sciences. La culture secrète les objets à mesurer, les conditions d'observation, les modes d'organisation de la statistique, le rôle de l'information quantitative dans les sciences, dans les procédures de prise de décision, dans la vie démocratique des nations. La technologie suggère des méthodes de collecte, de traitement, d'analyse et de diffusion. Réciproquement, la statistique et particulièrement la statistique officielle propose des grilles de décryptage de la réalité économique et sociale, des représentations qui conditionnent partiellement certaines décisions et structurent certaines actions. C'est de ce jeu subtil entre la statistique, la culture dans laquelle elle s'insère et les technologies sur lesquelles elle s'appuie que traite cet article. Il se propose plus particulièrement de se pencher sur la manière dont la statistique dite officielle ou publique risque d'être affectée par les changements culturels, administratifs et technologiques des années à venir. Il a donc un caractère prospectif et partant échappe sûrement au cadre scientifique habituel. Il présente des thèmes de réflexion, des intuitions sur les environnements culturel et technologique futurs alimentées à la fois par des expériences réalisées dans le cadre de projets communautaires de recherche en statistique (DOSES et DOSIS) gérés par Eurostat, l'office statistique des Communautés européennes et de réflexions et d'observations personnelles. Il s'inspire fortement d'une contribution que j'ai été invité à présenter lors de la 50ème session de l'institut international de statistique à Beijing en 1995.

L'article, après cette introduction, esquisse les relations entre la statistique et la culture d'une part, et la technologie d'autre part. Je tente ensuite d'identifier quelques évolutions probables du contexte administratif, culturel et technologique. J'analyse enfin l'impact des ces changements sur les principales opérations de la statistique publique: la conception des systèmes d'information, la collecte des données, leur traitement et leur stockage, l'analyse et le type d'utilisation des statistiques. Certaines leçons de développements attendus ou déjà en émergence sont tirées; des tendances antagonistes sont quelquefois mises en évidence.

2. Statistique et culture

La statistique s'est toujours mariée avec les différents siècles qu'elle a traversés. Dans l'antiquité, les dirigeants veulent dénombrer leurs sujets, comptabiliser leurs richesses et la statistique est une statistique de comptage à partir de recensements. Peu après la renaissance, lorsque les astronomes obtiennent plusieurs mesures d'un même phénomène, ils éprouvent le besoin de définir une mesure unique qui minimise les erreurs et la statistique leur propose des moyennes: Tycho Brahé a recours à la moyenne arithmétique à la fin du XVIème siècle. Il y a 3 siècles, des préoccupations financières amènent à s'interroger sur des variations de prix et le concept d'indices apparaît.

Parallèlement, la notion de probabilité fait son chemin. Les premières formulations font référence aux jeux de hasard largement pratiqués au XVIIème siècle: "Combien faut-il de

coups au moins pour que je puisse parier avec avantage que, après avoir joué ces coups, on aura obtenu un double six?" demande le chevalier de Méré à Pascal. Les mathématiques se formalisent et la notion de distribution se développe durant la deuxième moitié du XVIIIème siècle grâce aux travaux de Lagrange, Laplace, Gaus, pour en citer quelques uns. Ce n'est qu'au début du XIXème siècle que la statistique paraît quitter les sciences physiques et la gestion des Etats pour s'intéresser à la réalité sociale. Quételet publie en 1835 un livre sur la physique sociale qui se veut une étude de l'homme basée sur les probabilités; Pascal parlait seulement de géométrie du hasard en 1654 ...

La statistique adapte son objet et ses méthodes aux thèmes du jour. Et cette tendance n'appartient pas qu'au passé. La mesure de l'emploi, du chômage, des variations de prix, de la situation économique des entreprises s'enracine dans les dispositions administratives et comptables existantes. La mesure des aspects immatériels, de l'état "informationnel" (et non plus financier) des entreprises, comme certains l'ont appelé, reste difficile car peu couvert par la fiscalité, les pratiques juridiques et administratives. L'étude de la pauvreté est tributaire de la perception que la société nous en donne. Le statisticien officiel après avoir mesuré des nombres de têtes, réconcilié les mesures des astronomes, suivi des prix et fourni aux Etats les données dont il a besoin dans sa gestion, se penche aujourd'hui sur la pauvreté, la société duale et l'innovation, problèmes particulièrement préoccupants en cette fin du XXème siècle. Il ne paraît toujours pas s'affranchir de sa culture. Et pourtant, la relation n'est peut-être pas à sens unique. Les statisticiens à travers les représentations qu'ils construisent contraignent l'action politique, lui donnent des références, des instruments d'évaluation. Ne devrait-on pas en tirer des conclusions et quelque fois anticiper la demande en proposant à travers nos statistiques des descriptions de la réalité plus indépendantes du contexte? Une vue prescriptive en quelque sorte pour enrichir l'action économique, sociale et politique.

3. Statistique et technologie

Les influences croisées des technologies et de la statistique ne sont plus à souligner. Dans ce domaine également une rétrospective historique est éloquente. Des cartes perforées utilisées par Hollerith pour traiter les recensements américains à la fin du XIXème siècle aux ordinateurs parallèles, la statistique s'est émancipée au contact des nouvelles technologies. Comment imaginer le traitement des enquêtes actuelles sans ordinateur, comment effectuer des analyses factorielles, des classifications, étudier des structures multidimensionnelles au moyen de modèles de type LISREL sans des outils de calcul puissants et performants? L'explosion ces dernières années des technologies de l'information et des communications a eu et continuera d'avoir des impacts déterminants sur la statistique, discipline dont une des raisons d'être est justement de générer de l'information. L'intelligence artificielle, la télématique nous obligent à repenser nos méthodes, à les améliorer, à les généraliser. Et que nous réservent les années à venir?

4. Les évolutions culturelles, politiques et technologiques

Si la statistique dépend étroitement, comme suggéré dans cet article, de son environnement, nous ne pourrons prédire son évolution qu'en cherchant à anticiper la manière dont le climat administratif, culturel et technologique va se développer. Opération météorologique périlleuse! Mentionnons rapidement quelques tendances qui risquent

d'affecter le métier du statisticien; ceci sera repris plus systématiquement dans le paragraphe suivant sur l'impact sur la statistique.

- Un souci croissant de protection de la vie privée provoque l'éclosion de dispositifs juridiques appropriés, de comités divers.

- Le harcèlement administratif amène les redevables d'information à "bouder" certaines enquêtes, la notion même de recensement est remise en question.

- Les sondages aux moments critiques de la vie politique se multiplient; de plus en plus on gouverne en consultant, non plus seulement les parlements, mais également l'opinion publique et notre société est particulièrement friande de slogans, de raccourcis, de représentations schématiques que la statistique peut alimenter.

- La statistique se libéralise, des enquêtes font l'objet d'appels d'offres, les instituts nationaux de statistique sont mis en concurrence. Un véritable marché de l'information économique et sociale est en développement.

- Le paysage politique évolue; des tendances à la décentralisation qui provoquent un regain d'autonomie des autorités régionales et locales coexistent avec une communautarisation de certaines compétences, une centralisation de certaines activités; les exigences de comparabilité, d'harmonisation voire de consolidation transfrontalière augmentent (calculs d'agrégats européens à comparer avec les Etats-Unis, intérêt pour les groupes etc.).

En matière technologique, les évolutions sont peut-être plus faciles à anticiper, à court terme du moins. La "boule de cristal" à utiliser sont des enquêtes de type Delphi (auprès d'experts) organisées dans plusieurs pays, mais originellement au Japon, (depuis 1970) sur les technologies émergentes. Le taux de réalisation à 20 ans des prévisions des experts est en général assez faible; il paraît être de l'ordre de 25%. Dans le domaine des technologies de l'information ce taux est cependant substantiellement plus élevé. Les prévisions faites en 1971 au Japon se sont ainsi réalisées à 50% dans ce domaine au cours des 20 années suivantes et partiellement réalisées dans 80% des cas. Ceci nous invite à la modestie et à la prudence. Que nous annoncent les dernières études prospectives effectuées début des années 1990? L'apparition de techniques de développement rapide et de vérification de logiciels, de bases de logiciels qui puissent être réutilisés, de nouvelles techniques de programmation ne faisant plus appel à des langages, l'apparition d'ordinateurs qui puissent traiter des informations peu précises en utilisant le sens commun. Et tout ceci pour les années 2005 en moyenne (voir par exemple "Deutscher Delphi-Bericht zur Entwicklung von Wissenschaft und Technik" publié à Bonn en 1993 par Bundesministerium für Forschung und Technologie)

5. Impact des évolutions attendues sur la statistique.

Dans ce qui suit, l'impact sur les principales opérations de la statistique publique, des évolutions esquissées ci-dessus est analysé.

5.1. La conception des systèmes d'information

De l'information statistique au service statistique

Comme déjà mentionné, il apparaît évident que le futur sera caractérisé par plus de données et par plus d'opérateurs "statistiques " sur le marché, d'où très clairement pour le statisticien public une concurrence accrue: concurrence entre les secteurs public et privé mais également à l'intérieur de ces secteurs.

(Voir par exemple le nombre d'opérateurs mettant à disposition des répertoires d'entreprises.)

Ceci amènera probablement le statisticien à chercher à mieux vendre ses produits et à penser son métier plus en termes de services à rendre qu'en termes de données à fournir: les systèmes d'information statistique deviendront des services d'information statistique.

(Voir par exemple le développement de systèmes d'accès aux données du type EISI - développé dans le cadre du programme DOSES - qui partent de questions très générales de l'utilisateur et cherchent à leur répondre à partir de données existantes (Drappier, 1994 ou actes du séminaire NTTS tenu à Bonn en 1992).)

Ceci signifie entre autres la nécessité d'une meilleure documentation des données, un recours plus systématique à l'estimation, au redressement, à l'incorporation d'analyses dans les données fournies, à une étude de l'impact des résultats établis sur les décisions à prendre. Ceci signifie également le développement, comme on le voit de plus en plus souvent, de partenariats: la statistique officielle fournit des descriptifs de populations, incorpore des résultats établis par ailleurs, fournit une assistance méthodologique

De l'universel au sur mesure

Les systèmes seront sûrement plus diversifiés et plus ciblés que dans le passé. Une statistique sans ordinateurs puissants ne pouvait se payer le luxe de construire des systèmes personnalisés: des nomenclatures plus ou moins universelles avaient été établies et l'information était agrégée en conséquence. Les nouvelles technologies devraient nous permettre de stocker l'information élémentaire (par exemple un descriptif d'activité en langage naturel donné par un chef d'entreprise ou une réponse donnée à une question ouverte lors d'une enquête) et de ne coder cette information qu'au moment de l'utilisation en fonction des objectifs que l'on poursuit: des nomenclatures ad-hoc seront construites; elles ne seront pas le fruit unique d'une analyse a priori et souvent subjective des phénomènes à classer mais aussi le produit d'analyses a posteriori basées sur des proximités mesurées dans des espaces déterminés.

(Voir le développement de systèmes de classification automatique comme Synapse, le système développé par la société "inférence", ou le prototype construit dans le cadre de DOSES (actes du séminaire tenu en 1993 à Luxembourg sur les métadonnées).)

Le souci d'intégration des données, souci qui va croître avec la multiplication des sources d'information, amènera à développer des tables de correspondance entre ces nomenclatures.

De la donnée harmonisée à la métadonnée harmonisée

Les données devront être largement documentées, l'utilisation étant de plus en plus déconnectée de la collecte proprement dite, l'utilisateur étant de plus en plus éloigné du producteur (phénomène dû entre autres aux possibilités offertes par les réseaux télématiques); d'où une nécessité accrue, pour garantir un service de qualité, de normaliser cette documentation, à l'intérieur d'une application de la collecte à la diffusion, mais également entre applications.

(Des logiciels permettant de normaliser les métadonnées dans toute la chaîne statistique sont en cours de développement: famille de logiciels Blaise développée par le CBS aux Pays Bas, logiciel EMMA développé par la firme World Systems. Des projets internationaux comme le projet DSIS (Distributed Statistical Information Services) se proposent de développer des environnements dits de référence, harmonisés, c'est-à-dire entre autres documentés de manière standard (actes du séminaire NTTS tenu à Bonn en 1992).)

Du calcul sur les données au calcul sur les métadonnées

Cette information sur la donnée, souvent appelée métadonnée (terme un peu vague mais dont l'utilisation croissante est révélatrice de cette préoccupation émergeante), ne pourra être uniquement documentaire: elle devra non seulement qualifier la donnée, bien entendu, mais également suivre cette donnée lors de son traitement, d'où la nécessité de définir des algèbres de métadonnées appropriées, voire guider son traitement pour permettre des automatisations de certaines opérations. Des nombreux travaux expérimentaux ont été entrepris sur les nomenclatures: des langages et des logiciels sont en cours de développement (actes du séminaire tenu en 1993 à Luxembourg sur les métadonnées).

(Le traitement des drapeaux (flags) est particulièrement illustratif à cet égard. Que devient une annotation du type "valeur estimée" dans un calcul? Des travaux sont en cours à Eurostat sur ce sujet.)

De la mesure de l'erreur aléatoire à la mesure de l'incertitude

La banalisation de la donnée publique qui sera mise à disposition par de nombreux opérateurs sur le marché va amener le statisticien à cultiver ses spécificités: la production d'estimateurs de caractéristiques de populations et des erreurs qui les affectent. Le statisticien est le professionnel de l'erreur et la statistique est souvent qualifiée comme étant la première des sciences inexactes. Cette réputation a ses exigences! Qui peut mesurer précisément à l'heure actuelle l'erreur qui affecte l'estimation d'un PNB national? Comment se combinent les différentes sources d'erreurs qui sur toute l'histoire d'une donnée en affectent la valeur? Des mesures indépendantes des phénomènes permettent quelquefois de fournir une idée de la précision des estimations obtenues mais sans en chiffrer l'importance. Ces questions que l'utilisateur se posera de manière accrue lorsqu'il sera confronté à différentes sources d'information lui disant des choses différentes sur un même sujet, sont clairement de la compétence du statisticien. Il devra développer des mesures de l'erreur qui vont au delà de la modélisation du hasard, en cherchant peut-être du côté de la physique ou de l'intelligence artificielle des nouvelles métaphores pour fonder de nouvelles théories. Les travaux poursuivis dans le cadre des systèmes experts seront sûrement d'utiles sources d'inspiration (Cohen, 1985).

D'un monopole à la qualité comme argument de vente

De manière plus générale encore, dans un marché encombré par de nombreux acteurs, la notion de qualité va jouer un rôle accru; la statistique publique risque de perdre une partie de ses monopoles, elle devra combattre en garantissant une qualité de service imbattable. Elle devra tirer un parti maximum de son indépendance, de son objectivité. Cette exigence n'est pas le seul produit d'une compétition accrue entre les opérateurs statistiques, mais aussi la conséquence d'un impact croissant de la valeur de certaines statistiques sur la vie économique et sociale. Je reviendrai sur ce point essentiel dans la partie consacrée à l'utilisation des données. Le contrôle de qualité va donc imprégner l'ensemble des opérations statistiques, ses résultats seront intégrés à la donnée, et il fera dès le départ partie des systèmes d'information. Mais la qualité d'un service, ce n'est pas uniquement le niveau de fiabilité d'une information mais également les conditions dans lesquelles elle est délivrée. Différents types de services correspondant à différentes contraintes de précision, de délais devront être inventés. Les services postaux ont du développer des services de type DHL pour répondre à des besoins nouveaux et à une certaine concurrence. La statistique risque dans le futur d'être confronté à des besoins similaires.

(De nombreuses institutions ont développé par exemple des chartes de qualité et le concept de qualité totale a donné et donne lieu à de nombreux développements dans certains offices de statistique nationaux.)

5.2. La collecte des données

Les évolutions attendues (et pour certaines d'entre elles déjà observées) de la collecte des données résultent de différents effets des nouvelles technologies :

- méthodes de collecte et transmission automatiques de l'information via des techniques de type OCR (Optical Character Recognition), EDI (Echange de Données Informatisées); les données deviennent plus disponibles et se normalisent progressivement;

- complexification de la vie économique et par conséquent besoin accru d'informations d'où un constat d'harcèlement statistique des redevables de l'information;

- préoccupation croissante, face à une société de l'information qu'on imagine aisément inquisitrice, du respect de la confidentialité des données.

Examinons séparément l'impact probable de ces différents effets sur la collecte des données.

Les techniques OCR permettent d'automatiser des tâches fastidieuses de la statistique mais ont peu d'impact sur les méthodes utilisées. L'EDI, par contre, est, à mon avis, moins innocent et ce pour différentes raisons.

Premièrement il va déboucher sur une standardisation accrue des échanges avec les redevables d'information, d'où normalisation des messages mais aussi dans le sillage, des éléments d'information requis, y compris des métadonnées.

(Le volume de travail réalisé ces dernières années par le groupe MD6 dans le cadre de la norme EDIFACT est révélateur de l'importance de cette préoccupation au sein de la collectivité statistique. La nécessité d'une normalisation des métadonnées encapsulées dans le message apparaît également dans les travaux européens sur l'EDI comme SERT (Statistiques d'entreprises et réseaux télématiques).)

De la donnée statistique à la donnée utilisable à des fins statistiques

Deuxièmement, il va obliger les différents collecteurs d'informations à mieux se coordonner puisqu'ils utiliseront des moyens de transmission identiques: l'information collectée ne sera plus statistique mais pour partie au moins utilisable à des fins statistiques et dérivée automatiquement vers les services statistiques compétents; les concepts utilisés par les statisticiens vont donc devoir s'homogénéiser au sein des agences statistiques et se rapprocher des concepts utilisés par d'autres administrations (en matière de comptabilité par exemple).

(Les efforts faits dans des pays comme la France et les Pays Bas pour rapprocher les concepts comptables et statistiques, la mise en place du système Intrastat pour la collecte d'informations sur les échanges de biens et de services à l'intérieur de l'Union Européenne qui suit la fiscalité, l'utilisation de nomenclatures statistiques par des administrations fiscales, l'utilisation intensive de répertoires administratifs ou juridiques dans la construction des répertoires d'entreprises statistiques sont quelques exemples de cette tendance.)

D'un mode de collecte localisé dans le temps à un mode de collecte en continu

Troisièmement, les modes de collecte et les notions d'enquête risquent de devoir être repensés: les données seront fournies au fur et à mesure de leur disponibilité via des échanges entre différents systèmes d'informations; aux enquêtes localisées dans le temps succéderont des transmissions en "continu" de données venant des redevables d'information ou d'intermédiaires; les données statistiques dans leur collecte s'assimileront plus à des flux qu'à des stocks, d'où impact possible sur les théories d'échantillonnage par exemple.

De la collecte décentralisée aux grosses centrales de collecte

Enfin, le mode d'organisation de la statistique longtemps calqué sur les modes de collecte risque également d'être affecté. L'organisation d'enquêtes nécessite la création de centres locaux pour collecter les données qui sont ensuite traitées à différents niveaux avant d'être centralisées à l'étage national, régional ou mondial. Avec des techniques de collecte de type télédétection par exemple, la collecte et le traitement primaire des données appellent déjà une certaine centralisation qui pourrait avoir une incidence sur la répartition des rôles des différentes institutions impliquées dans ces opérations. Certaines données vont se concentrer de manière naturelle en certains endroits (banques centrales, offices de brevet, centres de vente, ...) auxquels le statisticien cherchera à avoir accès.

De l'enquête à la donnée administrative

La charge accrue des sujets enquêtés est un résultat indirect de la société de l'information et des nouvelles technologies qui a et aura de plus en plus un impact important sur les techniques de collecte de l'information. Le statisticien devra tirer parti de données existantes collectées souvent à des fins non statistiques provenant soit de sources administratives (ce qui se fait déjà dans de nombreux pays) ou de sources automatiques, pour utiliser une terminologie proposée par M. Volle (Volle, 1994).

L'exploitation des données administratives pose des problèmes spécifiques: les concepts sont divers, les données ont des degrés de fiabilité discutables dans certains cas, les populations couvertes ne coïncident pas toujours avec les populations que le statisticien veut étudier. Pour répondre à ces problèmes, un rhabillage des données est nécessaire, des contrôles, des cadrages, des appariements d'informations . Le statisticien officiel est en train d'acquérir un certain savoir faire en la matière.

(On peut mentionner à titre d'exemples, l'utilisation faite d'informations collectées par des centrales de bilan, et le remplacement dans certains pays de recensements par des sources administratives.)

Cette utilisation de sources administratives à différentes fins risque également d'avoir des impacts organisationnels et légaux : nécessité d'une meilleure coordination entre les différentes administrations, développement de cadres juridiques et administratifs pour réguler les échanges. Actuellement, les lois statistiques ne permettent pas souvent aux autorités statistiques d'accéder aux fichiers administratifs, des cloisonnements forts existent entre la fiscalité, la sécurité sociale, la justice, ... et la statistique. Le statisticien par sa neutralité et sa fonction multi sectorielle pourrait avoir un rôle central à jouer dans cette indispensable coordination.

De l'enquête à la source automatique

Des sources d'information nouvelles fournissent des données automatiquement: télédétection, échanges téléphoniques, énergétiques, caisses enregistreuses, gestion de mouvements sur les comptes bancaires, cartes de crédit, appareillages spécifiques pour mesurer par exemple le succès rencontré par certaines émissions de télévision, résultats de simulations, ... Le statisticien ne peut s'en désintéresser. De nouveau, l'utilisation de ces sources pose des problèmes spécifiques: granularité de l'information dont le niveau de détail dépasse ce que cherche la statistique, volume et fréquence des informations, caractère confidentiel des données, unités d'observations et modes d'identification (codes barre par exemple) différents. Ceci amènera le statisticien à repenser ses populations, éventuellement à définir des objets plus complexes (ensembles hiérarchiques ou en réseau comme un groupe d'entreprises, un tissu urbain, un comportement en matière de transports), à réinventer des indicateurs (basés par exemple sur des échanges téléphoniques pour chiffrer la santé d'une unité économique) qui tireront parti d'informations disponibles à coût réduit, à offrir des garanties visibles et encore plus convaincantes de la protection des données reçues et de leur non divulgation.

(Les dispositifs mis en place pour mesurer les taux d'observation de certaines émissions télévisées s'inscrivent par exemple dans ces évolutions.)

De la donnée publique à l'information confidentielle

Ce dernier point a une portée qui dépasse largement le cadre de l'utilisation de données automatiques à des fins statistiques (et le cadre de cet article, du reste) et son influence sur le métier statistique est considérable. L'histoire et certains fantasmes nous ont appris à redouter une centralisation excessive des informations. Il paraît exister une volonté individuelle, quelquefois un peu irrationnelle, à garder un jardin secret qui échappe à la juridiction des pouvoirs publics en général. Nos démocraties se portent garantes du respect de cette volonté de confidentialité et la statistique publique, tout en assurant la transparence nécessaire d'un certain nombre de phénomènes socio-économiques, doit en tenir compte. Pour ce faire, différentes techniques sont mises en oeuvre :

- questionnement introduisant des éléments aléatoires dont la distribution est connue de manière à perturber les réponses individuelles tout en gardant la possibilité de dériver certaines caractéristiques de population en épurant les résultats obtenus des aléas introduits;

- anonymisation des données;

- traitement particulier des informations confidentielles par recodification, arrondis, perturbation,

- génération de populations artificielles possédant des caractéristiques similaires aux populations réelles;

- protection des données via des environnements sécurisés, des techniques d'encriptage, des contrôles d'accès.

(De nombreuses conférences ont été organisées sur ce sujet (voir par exemple la conférence de Luxembourg de décembre 1994), des revues lui ont consacré des numéros, des logiciels génériques sont en développement.)

5.3. Le traitement et le stockage des données

Souligner l'impact de l'informatique sur les techniques de traitement statistique est devenu un lieu commun. Les moyens mis à disposition deviennent de plus en plus performants et autorisent des traitements de plus en plus sophistiqués sur lesquels je reviendrai dans la partie consacrée à l'analyse des données.

De la base de données centrale aux bases réparties

Un des faits majeurs auquel le statisticien devra faire face en matière de traitement et de stockage est sûrement la morcellisation des sources d'information, des bases de données. La démocratisation informatique a favorisé l'éclosion de bases de données locales, l'éparpillement des sources. La donnée devient de plus en plus immatérielle, circule sur des réseaux et les systèmes d'information doivent se concevoir plus comme des systèmes nerveux où différents agents collaborent que comme des systèmes centraux où les informations sont stockées. Ceci pose des problèmes matériels de câblage et communication, mais surtout des problèmes de dialogue et d'harmonisation. Plusieurs institutions qui mettent des données en commun doivent parler une même langue statistique, doivent synchroniser leurs opérations. La nécessité de la documentation des données a déjà été soulignée dans le paragraphe consacré aux systèmes d'information; le partage de ressources rend également cette documentation indispensable. Au même titre qu'un consommateur dans un hyper marché souhaite comparer différentes marques pour un même produit, l'utilisateur de statistiques doit savoir d'où viennent les informations qui lui sont accessibles, pouvoir les localiser dans le temps, les comparer. Les progrès réalisés par la bureautique communicante qui tendent à gommer de plus en plus l'existence de différents machines localisés dans différents sites, accroissent cette nécessité.

(Le projet européen DSIS déjà mentionné a pour vocation essentielle la mise en réseau de différentes bases nationales existantes. Des travaux organisés dans le cadre du programme de recherche ENS (European Nervous Systems), par exemple, permettent de relier des répertoires nationaux gérés par des chambres de commerce.)

Du traitement séquentiel au traitement en parallèle

Les possibilités accrues de stockage nécessitent des réécritures de certains algorithmes statistiques et le développement de nouvelles méthodes de traitement. Il s'agit ici d'adapter des méthodes connues à des grands volumes de données et à mieux tirer parti des caractéristiques des outils de calcul existants. L'apparition d'ordinateurs opérant en parallèle risque d'avoir un impact certain sur les algorithmes statistiques. Diverses expériences ont été réalisées en statistique officielle entre autres dans le domaine de la codification automatique des activités et des occupations (Creecy, 1992).

Du quoi au comment

L'automatisation du traitement exige sa formalisation. Quelles sont les heuristiques à utiliser pour faire de la codification automatique (Drappier, 1994)? Comment estimer une donnée manquante? Comment corriger une donnée des variations saisonnières? Les traitements possibles vont bien au delà d'une simple manipulation numérique; ils exigent le développement de stratégies complexes combinant le calcul simple à l'utilisation de règles, à la recherche d'informations complémentaires etc. (Hand, 1986). L'intelligence artificielle nous apporte en la matière un savoir faire précieux qui ne peut être que plus utilisé dans le futur: stratégies de résolution de problèmes, apprentissage, reconnaissance de formes... Le mérite de ces évolutions déjà observées n'est pas uniquement de libérer le statisticien d'un certain nombre de tâches mais également comme mentionné en introduction de mettre l'accent sur le "comment". L'objet d'étude se déplace de la formalisation de structures à la formalisation de processus. Ces nouveaux accents permettent de débattre des stratégies qui auront, pour les besoins d'automatisation, été explicitées. Le problème de validation de ces stratégies n'est cependant pas à sous estimer. Quand peut-on affirmer qu'une méthode est bonne ou meilleure qu'une autre? Dès qu'une méthode est encapsulée dans un programme ou un logiciel, elle acquiert un statut qui pourrait tromper l'utilisateur. On risque en effet d'induire une confusion entre les éléments "mathématiques" d'un processus qui sont en quelque sorte indiscutables et "prouvés" et des éléments plus subjectifs qui résultent de choix faits par certains experts. Imaginez un traitement médical où on ne puisse pas faire la part entre ce que donne le bilan sanguin et le diagnostic qui en résulte...

(L'expérience acquise dans DOSES est à cet égard instructive (Drappier, 1994). Le projet ALIEN développé par J.L. Roos de l'INSEE, par exemple, permet à partir d'un ensemble d'indicateurs économiques numériques de générer de manière automatique des commentaires sur la conjoncture. Ceci ne peut se faire que via une formulation des règles heuristiques appliquées par les conjoncturistes. Une règle explicite est ouverte au débat, à l'amélioration, à l'échange.)

5.4. L'analyse

Les défis identifiés jusqu'à présent concernent plus le statisticien officiel que le statisticien "universitaire". Dans ce paragraphe je vais m'attacher à évoquer des progrès que les nouvelles technologies vont offrir à la statistique en temps que discipline académique, à son objet et à ses méthodes. Les nouvelles technologies fournissent en effet des possibilités de manipulation des informations qui pourraient élargir considérablement l'éventail des méthodes mais également le champ de la statistique.

Du modèle théorique à l'estimation

La statistique a longtemps été tributaire d'hypothèses sur les distributions des variables pour réaliser des tests, effectuer des estimations. L'informatique et les simulations qu'elle rend possibles a ouvert le champ à des nouvelles méthodes de calcul des erreurs (méthode de type bootstrap) et par conséquent à plus d'audace dans le choix des plans d'échantillon et des statistiques à considérer.

De la référence temporelle à la référence spatio-temporelle

En statistique officielle, le temps joue un rôle privilégié. Il dessine des évolutions, des tendances qui permettent de mieux évaluer, de mieux comprendre les phénomènes observés. L'analyse des séries chronologiques occupe du reste une place importante dans le métier du statisticien. De manière a priori paradoxale, la référence spatiale est beaucoup plus rare; les observations sont attachées à des agrégats géographiques rarement à des points précis et peu d'utilisation paraît être faite de cette localisation. Ceci est du à différentes raisons: peu d'importance accordée par la théorie économique aux éléments spatiaux, utilisation d'échantillons ne permettant pas une exploitation locale pour des raisons de fiabilité des résultats, mode d'enregistrement des données. L'avènement des systèmes d'informations géographiques et la présence de données finement géo-référencées sont en train de changer cet état de fait. Des nouvelles méthodes tirant parti des coordonnées spatiales, des modes de présentation, de stockage des données spatiales se développent. La référence spatiale apparaît comme facteur d'intégration de données d'origine diverses, les informations statistiques et géographiques se marient (Heath, 1994).

De la donnée numérique à la donnée symbolique

Le paradigme classique de la discipline est l'étude de la distribution d'une ou d'un ensemble de caractéristiques sur une population d'unités données; les caractéristiques sont nominales, ordinales ou quantitatives, pour faire bref. En fait, cette représentation de la réalité nécessite souvent d'opérer des simplifications dans les données, de faire fi de certaines informations: on oubliera lorsqu'on étudie des régions caractérisées par certaines variables socio-économiques la proximité entre ces régions, on caractérisera le comportement innovateur d'une entreprise à partir de quelques variables choisies en oubliant la composante temporelle et les phénomènes d'interaction, des accidents ne seront pas représentés par des scénarios types mais par un vecteur de variables etc. Ces simplifications deviennent de moins en moins nécessaires car les outils disponibles ont forcé le développement de nouveaux paradigmes. Les objets qui intéressent l'utilisateur sont souvent des objets complexes (des groupes d'entreprises, des comportements, des structures), symboliques (des classes d'objets, des prototypes) caractérisés par des variables structurées (intervalles, nomenclatures hiérarchiques, variables dont les modalités sont dépendantes, ...) et des régularités traduisibles en règles, en intervalles de variation... Aux notions de classement d'unités se substituent des notions de description simple d'objets complexes au moyen de prédicats (Diday, 1993). Le statisticien peut maintenant élargir la gamme d'objets à représenter, la gamme des mesures à utiliser et des opérateurs à définir.

De la donnée à l'information

Il peut également se rapprocher dans ses analyses du langage et des concepts utilisés par les utilisateurs: les objets symboliques interviendront en entrée et en sortie de ses travaux. Les nouvelles technologies posent à cet égard des défis fascinants qui dans les années à venir

élargiront le champ de compétence du statisticien de la donnée numérique à la donnée structurée. Ces progrès ne laissent pas le statisticien européen que je suis indifférent; ils permettent de rapprocher le langage du statisticien de celui de l'utilisateur final et ce faisant de laisser petit à petit moins de place à la subjectivité (et par conséquent la non comparabilité) dans l'interprétation des résultats. Imaginez le résultat d'une même analyse factorielle par exemple interprété par des utilisateurs de différentes nationalités ou de différents secteurs. Quelle signification attribuer aux axes? Du résultat mathématique à son habillage conceptuel, que de chemin! Les techniques d'analyse symbolique en cours de développement devraient faire reculer les frontières de l'interprétation libre en proposant des descriptifs structurés de résultats dans des langages qui se rapprochent de plus en plus du langage de l'utilisateur: le résultat doit signifier plus d'harmonisation et par conséquent une meilleure comparabilité du message transmis.

(Des nombreux travaux entrepris en Intelligence Artificielle sur la formation de concept (Michalski, 1986) sont en fait proches de ce qui se fait en statistique; ils ne se limitent pas à un traitement numérique de l'information mais cherchent à construire des descriptifs en utilisant par exemple le calcul des prédicats.)

De la métaphore physique à la métaphore biologique

Les réseaux neuronaux et les algorithmes génétiques paraissent également susceptibles d'élargir l'éventail des techniques statistiques. Les métaphores biologiques se substituent aux métaphores physiques (centres de gravité, inerties ...). Les systèmes dits connexionnistes offrent des nouveaux modes de calcul et des nouveaux types de représentation. Ils permettent par exemple d'ajuster" des modèles non linéaires sur des séries chronologiques et de se passer d'une formalisation analytique a priori des relations à modéliser. Ils proposent des décompositions canoniques de certaines fonctions. En matière de représentation des résultats, les cartes de Kohonen généralisent les représentations classiques dans des espaces multidimensionnels, par des représentations sur des réseaux plus ou moins complexes (Varfis, 1992). Ils imposent des relations de proximité qui généralisent en quelque sorte ce qu'on obtient dans des espaces plus classiques.

(A. Varfis a ainsi comparé certaines méthodes classiques de prévision avec des méthodes empruntées aux réseaux neuronaux et les résultats obtenus au moyen de cartes de Kohonen aux résultats obtenus par des analyses factorielles; les perspectives offertes par ces nouvelles techniques paraissent prometteuses.)

De l'algorithme numérique à la formalisation de la démarche de l'expert

Les nouvelles technologies paraissent susceptibles de jeter des ponts entre la connaissance intuitive que l'utilisateur a du monde et les modèles mathématiques utilisés pour le décrire. A cet égard des travaux menés dans le cadre de DOSES, ont montré comment en matière de prévision on pouvait mieux confronter les modèles utilisés par les experts non statisticiens et ceux plus formalisés utilisés en statistique en utilisant par exemple des diagrammes d'influence (Talbot, 1992). De nouveau, le défi posé au statisticien est une utilisation pertinente des nouveaux modes de représentation et de manipulation des informations symboliques offerts par l'ordinateur.

5.5. L'utilisation des statistiques

La complexité croissante de la vie en société rend l'information de plus en plus précieuse. Cette réflexion banale appelle cependant des commentaires qui le sont peut-être moins :

- l'information est plus qu'un ensemble de données;
- elle constitue un bien qui a une valeur pour le chef d'entreprise, le chercheur, le citoyen, ...

De la collecte au raffinage

De manière un peu paradoxale, nous souffrons déjà d'un trop plein de données et ceci ira en s'accroissant. "Tout individu ou toute organisation recevra à l'état brut plus d'informations qu'il ne pourra en utiliser et il lui faudra trier et réduire cette information" (Lesourne, 1995). Le défi pour nous statisticiens est d'utiliser ces informations brutes (souvent des données numériques) pour en extraire des informations synthétiques qui permettent soit d'agir, soit d'enrichir nos connaissances et notre compréhension. Le rôle du statisticien officiel risque progressivement de se transformer de mineur, extracteur de matières premières (les statistiques de base), en raffineur, producteur d'informations. Les méthodes présentées au paragraphe précédant vont dans ce sens là. Le raffinement imposera sûrement un tri sévère des données, pour faire la part de ce qui est objectif, transformé, retraité, voire manipulé dans les informations qui circulent.

De la production de statistiques à l'arbitrage

La valeur économique de certaines informations statistiques n'est plus à démontrer; il suffit de voir l'impact de la publication d'une balance des paiements sur une économie nationale pour s'en persuader. Mais sa valeur ne se réduit pas à cela. Nous vivons dans une société friande d'images, de slogans, de raccourcis. La statistique est un instrument privilégié pour construire ces représentations. Le rôle des sondages ou des statistiques dans une campagne électorale est particulièrement illustratif à cet égard. On ne gouverne pas uniquement à travers des parlements mais également et de plus en plus à travers l'image de l'opinion et de l'économie reflétée à travers des médias et souvent à partir de données statistiques. Mais les messages véhiculés peuvent être difficiles à comprendre, ambigus, trompeurs, voire contradictoires. La responsabilité du statisticien est grande; face à cette avalanche de chiffres il doit arbitrer, expliquer, éduquer. Le rôle de plus en plus stratégique joué par la statistique appelée à justifier des allocations budgétaires et des dépenses créera des tensions et renforcera cette nécessité d'une expertise statistique. De nouveau ces inflexions risquent d'avoir un impact substantiel sur le rôle du statisticien officiel. Il se transforme en témoin, instrument de validation, voire juge.

(Vouloir juger de l'impact des activités de recherche et développement à partir de données sur les brevets, d'indicateurs sur ce qui est publié, de références à des innovations dans des revues spécialisées, de balances des paiements technologiques ou d'enquêtes statistiques ad-hoc débouche sur des résultats différents, voire contradictoires. Le statisticien économiste est en position idéale pour analyser ces disparités et les expliquer aux preneurs de décision.)

De la production de statistiques à la formation

La statistique a une valeur sociale et fait partie intégrante des démocraties: si l'information a une valeur, cette valeur doit être équitablement répartie; tout citoyen doit

pouvoir accéder à des données sur l'état de son économie, doit pouvoir les comprendre pour juger d'une gestion. Ces données doivent être impartiales, fiables, pertinentes et compréhensives: le statisticien doit être le garant de ces qualités des statistiques publiques. Ce défi n'est pas nouveau mais il est rendu particulièrement aigu par l'explosion des moyens de communication et par l'accroissement des sources d'information. Pour y faire face, le statisticien devra se rapprocher de l'utilisateur, se faire pédagogue, et affiner ses outils.

(L'impact des sondages sur l'opinion publique, de données désaisonnalisées sur l'emploi, de données sur le commerce extérieur etc. nécessite une information et une formation du public et plus particulièrement de ceux qui sont amenés à utiliser ces statistiques.)

6. Conclusions

La plupart des évolutions évoquées dans cet article sont déjà en cours. Elles refaçonnent progressivement le travail du statisticien, sa formation, son environnement technique et les institutions. La vitesse d'adaptation de ces différents éléments n'est malheureusement pas la même. Il n'est pas impossible que les années à venir se caractérisent par des tensions entre d'un côté des utilisateurs et un environnement technologique avancés (autoroutes de l'information, systèmes experts, télétravail ...) et de l'autre des institutions et des méthodes en lente mutation. Certaines activités menées par des services statistiques nationaux ou internationaux ont pour ambition d'anticiper ces tensions; des travaux de recherche sont entrepris pour moderniser l'appareil statistique et mieux le préparer à affronter les défis posés par les nouvelles technologies. Eurostat dans le sillage du programme DOSES a lancé en 1995 le programme DOSIS (Development of Statistical Information Systems). Son objectif est entre autres de servir de catalyseur au niveau européen en réunissant services publiques, entreprises et chercheurs, fournisseurs et utilisateurs de données, autour de thèmes et d'enjeux communs; il veut ainsi favoriser l'émergence de nouveaux modèles, de nouvelles techniques, de nouveaux logiciels. Il doit également contribuer à mieux répartir le savoir faire; il importe d'éviter que les technologies renforcent les disparités existantes: le progrès passe par la cohésion. C'est peut-être le plus important des défis.

Le futur n'est pas uniquement une vue de l'esprit, il n'est pas inéluctable, il s'élabore à partir de projets politiques, de travaux de recherche, de volonté individuelle; le prévoir, c'est déjà commencer à le construire.

REFERENCES

Bundesministerium für Forschung und Technologie (1993), *Deutscher Delphi-Bericht zur Entwicklung von Wissenschaft und Technik*, Bonn.

Cohen P.R (1985), *Heuristic reasoning about uncertainty: an artificial intelligence approach,* Pitman Advanced Pub. Program, Boston.

Creecy R. (1992), "Massively parallel computing and automated industry and occupation coding", Eurostat, *Proceeding of the seminar on new techniques and technologies for statistics*, Bonn.

Diday E. (1993), *Quelques aspects de l'analyse des données symboliques,* Rapport de recherche INRIA n° 1937, Paris.

Drappier J. (1994), *DOSES, its evaluation, its results, its future*, published by the Office for Official Publications of the European Communities, Luxembourg.

Drappier J. (1993), "Statistical Meta Information Systems", *Proceedings of the conference*, published by the Office for Official Publications of the European Communities, Luxembourg.

Eurostat (1992), *New Technologies and Techniques for Statistics, Proceedings of the conference*, published by the Office for Official Publications of the European Communities, Luxembourg.

Eurostat (1994), *Proceedings of the seminar on statistical confidentiality*, published by the Office for Official Publications of the European Communities, Luxembourg.

Hand D. (1986), "Patterns in statistical strategy", in W.A. Gale (editor), *Artificial Intelligence & Statistics,* Addison-Wesley Publishing Company.

Heath D.W. (1994), *A general view of GIS*, contribution présentée à la conférence SCORUS tenue à Helsinki.

Lesourne J. (1995), *Sigma*, published by the Office for Official Publications of the European Communities, Luxembourg.

Michalski R.S. (1986), "A theory and methodology of inductive learning", in R.S. Michalsky, J.G. Carbonell, and T.M. Mitchell (editors), *Machine learning: an Artificial Intelligence Approach,* Tioga Publishing Company, Palo Alto.

Talbot M. (1992), "Linking Informal Knowledge and Expertise to Forecasting Models", *Proceedings of the seminar on New Technologies and Techniques for Statistics,* in Eurostat (editor), published by the Office for Official Publications of the European Communities, Luxembourg.

Varfis A. and Versino C. (1992), "Clustering of socio-economic data with Kohonen Maps", *Neural Network World, 2, N° 6,* pp. 813-834

Volle M. (1994), *Traitement statistique des données collectées automatiquement*, rapport interne Eurostat, Luxembourg